Library Hi Tech

THC-DAT helps in reading a multi-topic document: results from a user-centered evaluation of a within-document analysis tool

Jing Chen Quan Lu Dan Wang Zeyuan Xu

Article information:

To cite this document:
Jing Chen Quan Lu Dan Wang Zeyuan Xu , (2016),"THC-DAT helps in reading a multi-topic document: results from a user-centered evaluation of a within-document analysis tool", Library Hi Tech, Vol. 34 Iss 4 pp. -
Permanent link to this document:
http://dx.doi.org/10.1108/LHT-07-2016-0081

THC-DAT Helps in Reading a Multi-topic Document: Results from a User-Centered Evaluation of a Within-Document Analysis tool

1. Introduction

With the rapid growth of electronic documents, there are more and more multi-topic documents, which requires users with high reading ability. Compared with single topic documents, multi-topic documents, including scientific articles, news stories and patents, may come naturally with complex hierarchical structure, involving more than one topic, meanwhile topics within separated paragraphs may have relationships (Chen, Wang & Lu, 2016; Tagarelli & Karypis, 2013). So analyzing the document from within-document perspective then showing the content in a organized way is helpful for users in understanding the multi-topic document. In respect to how to visualize and analyze these multi-topic documents, various tools have been proposed, such as TOPIC ISLANDS (Miller, Wong & Brewster, 1998), HINATA (Nishihara, Sato & Sunayama, 2011) and TopicNets (Gretarsson *et al.*, 2012). Unfortunately, most tools ignore the latent hierarchical structures and the context information within a document. Hence, we have proposed a new multi-topic document analysis tool THC-DAT (as acronym of Topic Hierarchy and Context based Document Analysis tool) which uses hierarchical Latent Dirichlet Allocation (hLDA) method to extract a topic hierarchy tree, and takes the context information into account to enable users to browse and analyze a document in a multi-grained, topic-oriented and context-based way (Chen *et al.*, 2016).

Text signal is introduced as the writing device that emphasizes aspects of a text content or structure without adding to the content of the text (Lorch, 1989). It attempts to pre-announce or emphasize content and reveal content relationship (Lemarié, Eyrolle & Cellier, 2006; R. Lorch, E. Lorch & Ritchey, 2001; Spyridakis, 1989), and can direct attention of readers during reading and help improve readers' ultimate comprehension about text information. Actually, the hierarchical topic tree extracted by THC-DAT is one kind of text signals. BOOKMARK such as the Table of Contents bookmarks in Adobe Reader, which provides a directory tree that integrates all levels of titles of a document, is the most common text signal and within-document analysis tool. In our previous work, we provided a comprehensive overview of approaches, interface and functions of THC-DAT. Subsequently, we conducted a case study to evaluate the tool, and qualitative analysis results that indicate the effectiveness of the tool were also discussed. In this paper, we conducted a comparative evaluation of THC-DAT with BOOKMARK to figure out whether THC-DAT enables users to browse, analyze and understand a multi-topic document more efficiently and effectively. With this intent, we investigated the two tools within a simulated work task situation, in which participants were asked to finish three kinds of tasks about a document, each tool was used to finish three tasks. On the basis of quantitative performance data and qualitative data derived from questionnaires, we assessed the comparative efficiency, effectiveness and user satisfaction of the tools.

The structure of the paper is as follows. Section 2 reviews related work. Section 3 introduces the interfaces of the tools used in the experiment. Research questions and hypotheses are provided in Section 4, the whole experiment design scheme is presented in Section 5. Section 6 shows the results of the experiment study and the discussion. Finally, some concluding remarks are offered in Section 7.

2. Related work

With the growing availability of electronic document in recent years, topic-based tools which reveal the topic structure in a long document with multiple topics are becoming a research hotspot. For example, TOPIC ISLANDS (Miller *et al.*, 1998) applied wavelet technology to extract topics. HINATA (Nishihara *et al.*, 2011) visualized topic-sentence relevance in documents by light and shadow. TopicNets (Gretarsson *et al.*, 2012) considered unique internal sequence of linear structures and analyzed the topic with statistical topic models. TIARA (Liu *et al.*, 2009) applied latent Dirichlet allocation (LDA)-based topic analysis to automatically derive topics and content evolution over time. Topic hypergraph (Wang *et al.*, 2013) analyzed topic structure for long documents divided into multiple segments, and extracted topic by LDA algorithm. In general, tools of this type are limited to analyze particular type of document and topics they extract neglect semantic relations.

Since these within-document analysis tools are ultimately used by users, usefulness and usability are main concerns of researchers, and user-centered evaluation becomes quite necessary. At present, the evaluation methods of within-document analysis tools mainly include control experiment and usability test .

Control experiment, which conducted in a simulated working environment, is the most common method. In a control experiment, the tool to be evaluated will be compared with a reference tool, with participants completing the same or similar task on the tools. A common tool was often selected as the baseline. For instance, Mizoguchi, Sakamoto & Igarashi (2013) proposed four types of overview scrollbars and compared them with the traditional scrollbar. Wu, He & Xu (2012) examined two relevance feedback techniques in interactive multilingual information access (MLIA), and regarded a basic interactive MLIA search without any relevance feedback as a benchmark system. Byrd (1999) compared an experimental system incorporating full visualization to a control system with no visualization, except for highlighting words in a single color. FindSkim, based on the ubiquitous Find-Command, served as a benchmark to compare with ProfileSkim based on relevance profiling in Harper, Koychev & Sun's study (2004). When it comes to task design, tasks are often devised to satisfy test assumptions or research purposes. For example, Mizoguchi *et al.* (2013) set all tasks to search and click objects on a vertically long document to investigate the performance of overview scrollbar. Whittaker *et al.* (2010) attempted to test a new user interface (SCAN) in local browsing by comparing different retrieval situations with relevance ranking, fact-finding and summarization tasks. Liu *et al.* (2009) designed three types of email analysis tasks to evaluate the effectiveness of TIARA in support of analysis tasks of different difficulty. Dang *et al.* (2012) proposed Nano Mapper to support users' search and analysis of nanotechnology developments and classified tasks based on the search functionality involved.

Usability test is used in user-centered interaction design to evaluate a system by testing it on users. It focuses on measuring a system's capacity to meet its intended purpose. Researchers usually evaluate the system from two perspectives: objective and subjective measures. Objective measures are derived from completion time, answer accuracy and log data to assess effectiveness and efficiency. Subjective measures are users' response or preference to the system, which is usually estimated by questionnaire. For example, Byrd (1999) made objective and subjective measurements to appraise the experimental tool. Harper *et al.* (2004) evaluated ProfileSkim by effectiveness, efficiency and user satisfaction. Hersh, Pentecost & Hickam (1996) compared two MEDLINE searching systems in aspects of answer accuracy, completion time, relevant articles retrieved and user satisfaction. Schwartz, Hash & Liebrock (2010) conducted a user study to evaluate visualizations with Focus+Context model, which measured quantitative results and qualitative responses.

To sum up, we conducted a control experiment to examine the reading functions of THC-DAT compared with BOOKMARK. Six question-answering tasks that can be classified into three categories were set based on the research purpose. Additionally, the efficiency, effectiveness and user satisfaction were regarded as the measurement indexes during the usability test.

3. Experimental tools

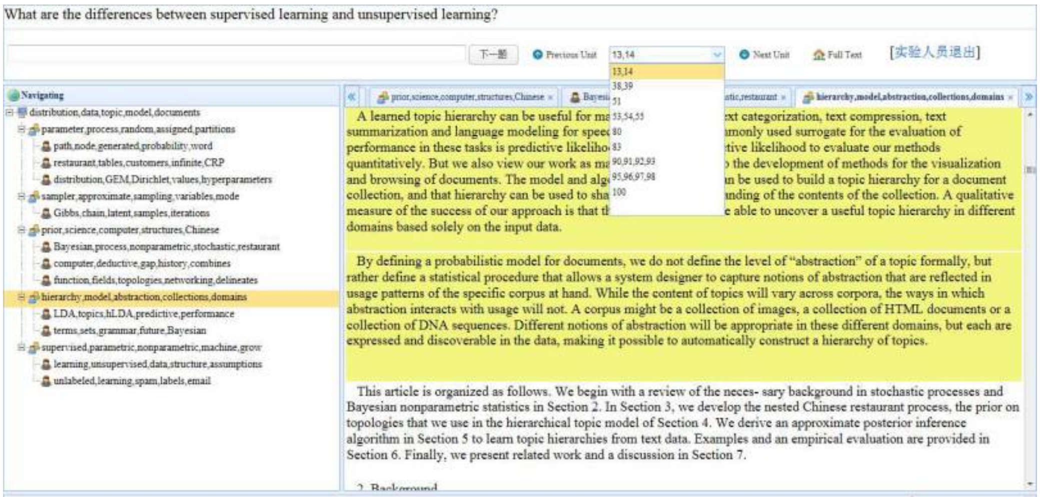
Two within-document analysis tools, THC-DAT and BOOKMARK, are used in the control experiment. Although the two tools are designed based on two different ideas, both interfaces are similar. Functions and operations of the two tools are described as follows.

3.1 THC-DAT

THC-DAT is designed based on topic hierarchy and context information. It applies hLDA method to visualize a topic hierarchy tree in a fine-grained way, in which users can search and browse according to their interested topics to obtain relevant paragraphs and analyze hierarchical document structure. Here, each node in the tree represents a topic with its terms and has a set of corresponding paragraphs representing contents of the topic. Furthermore, by context information, users can quickly grasp the distribution of topics in the document and find out the relationships between paragraphs and topics.

A screenshot of the THC-DAT tool is illustrated in Figure 1.

The leftmost interface is a hierarchical topic tree whose structure and five key topic terms of every node are generated by HLDA algorithm. The document is divided by the topic tree which aggregates paragraphs that express the same topic. Note that the topic tree in Figure 1 has a three-level structure. A root node can summarize the topic of the full-text abstractly, further being divided into five specific second-level nodes, with each one summarizing the topic of a part of text, then these sub-nodes will be divided into certain third-level nodes more specifically. Namely general topic corresponds to the root while specialized topic corresponds to the leaf. A new tab page will be generated on the right of the scrollable panel when clicking on any topic node at a time, the same tab can't appear twice but only refresh. Meanwhile paragraphs covered with the clicking topic node will be highlighted. The document is positioned in the display panel at the first paragraph under the topic node, and paragraphs may not continuous but scattered across the full document. All paragraph numbers under the topic node will be displayed in the drop-down box on the top, merging adjacent paragraphs. The document display panel can jump to the corresponding paragraph when users per click a paragraph number unit. Moreover users can navigate from the current paragraph to the previous (or next) paragraphs by "Previous Unit" and "Next Unit" buttons. In a word, with the topic tree, users can have a preliminary understanding of distribution and relationship of topics, through reading corresponding contents then compare different contents between topics.


下一题
Previous Unit
13,14
Next Unit
Full Text
[实验人员退出]
13,14
38,39
51
53,54,55
80
83
90,91,92,93
95,96,97,98
100
Navigating
distribution,data,topic,model,documents
parameter,process,random,assigned,partitions
path,node,generated,probability,word
restaurant,tables,customers,infinite,CRP
distribution,GEM,Dirichlet,values,hyperparameters
sampler,approximate,sampling,variables,mode
Gibbs,chain,latent,samples,iterations
prior,science,computer,structures,Chinese
Bayesian,process,nonparametric,stochastic,restaurant
computer,deductive,gap,history,combines
function,fields,topologies,networking,delineates
hierarchy,model,abstraction,collections,domains
LDA,topics,hLDA,predictive,performance
terms,sets,grammar,future,Bayesian
supervised,parametric,nonparametric,machine,grow
learning,unsupervised,data,structure,assumptions
unlabeled,learning,spam,labels,email
prior,science,computer,structures,Chinese ×
Bayesi
stic,restaurant ×
hierarchy,model,abstraction,collections,domains ×
2. Background

What are the differences between supervised learning and unsupervised learning?

A learned topic hierarchy can be useful for ma … xt categorization, text compression, text summarization and language modeling for speec … monly used surrogate for the evaluation of performance in these tasks is predictive likeliho … tive likelihood to evaluate our methods quantitatively. But we also view our work as ma … the development of methods for the visualization and browsing of documents. The model and alg … n be used to build a topic hierarchy for a document collection, and that hierarchy can be used to sha … nding of the contents of the collection. A qualitative measure of the success of our approach is that th … e able to uncover a useful topic hierarchy in different domains based solely on the input data.

By defining a probabilistic model for documents, we do not define the level of "abstraction" of a topic formally, but rather define a statistical procedure that allows a system designer to capture notions of abstraction that are reflected in usage patterns of the specific corpus at hand. While the content of topics will vary across corpora, the ways in which abstraction interacts with usage will not. A corpus might be a collection of images, a collection of HTML documents or a collection of DNA sequences. Different notions of abstraction will be appropriate in these different domains, but each are expressed and discoverable in the data, making it possible to automatically construct a hierarchy of topics.

This article is organized as follows. We begin with a review of the neces- sary background in stochastic processes and Bayesian nonparametric statistics in Section 2. In Section 3, we develop the nested Chinese restaurant process, the prior on topologies that we use in the hierarchical topic model of Section 4. We derive an approximate posterior inference algorithm in Section 5 to learn topic hierarchies from text data. Examples and an empirical evaluation are provided in Section 6. Finally, we present related work and a discussion in Section 7.

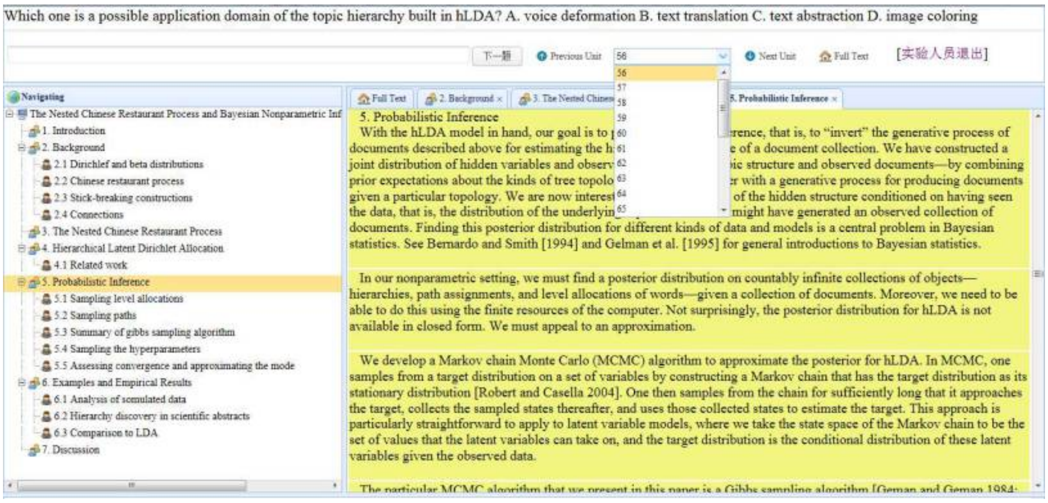
Figure 1 Screen shot for THC-DAT

3.2 BOOKMARK

Our benchmark system BOOKMARK is an essentially tagged PDF file. It is based on the table of contents which could reveal the hierarchical topic within the book (Cao & Wang, 2000; Yin, 2012; Liu, 2003). BOOKMARK has a mechanism combining contents and logical structures through inherent tags, such as chapter, section, tables etc. So it expresses the topic structure in a coarse-grained way. Since BOOKMARK is a ubiquitous tool, similar to the table of content of a book, and thus it can be understood easily.

A screenshot of BOOKMARK is illustrated in Figure 2.

A directory tree that extracts table of contents of the document is on the left, a new tab page can be generated on the right of the scrollable panel when clicking on any title at a time. Similar to the THC-DAT tool, paragraphs correspond to the title node will be highlighted, and all paragraph numbers under the title will display in order in the drop-down box. So with the tool, users can navigate title catalogs and read passages under a certain title, then compare different contents within different titles.


下一题
Previous Unit
56
Next Unit
Full Text
[实验人员退出]
56
57
58
59
60
61
62
63
64
65
Navigating
The Nested Chinese Restaurant Process and Bayesian Nonparametric Inf
1. Introduction
2. Background
2.1 Dirichlef and beta distributions
2.2 Chinese restaurant process
2.3 Stick-breaking constructions
2.4 Connections
3. The Nested Chinese Restaurant Process
4. Hierarchical Latent Dirichlet Allocation
4.1 Related work
5. Probabilistic Inference
5.1 Sampling level allocations
5.2 Sampling paths
5.3 Summary of gibbs sampling algorithm
5.4 Sampling the hyperparameters
5.5 Assessing convergence and approximating the mode
6. Examples and Empirical Results
6.1 Analysis of somulated data
6.2 Hierarchy discovery in scientific abstracts
6.3 Comparison to LDA
7. Discussion
Full Text
2. Background ×
3. The Nested Chines
5. Probabilistic Inference ×

Which one is a possible application domain of the topic hierarchy built in hLDA? A. voice deformation B. text translation C. text abstraction D. image coloring

5. Probabilistic Inference

With the hLDA model in hand, our goal is to p … rence, that is, to "invert" the generative process of documents described above for estimating the h … e of a document collection. We have constructed a joint distribution of hidden variables and observ … ic structure and observed documents—by combining prior expectations about the kinds of tree topolo … r with a generative process for producing documents given a particular topology. We are now interest … of the hidden structure conditioned on having seen the data, that is, the distribution of the underlyin … might have generated an observed collection of documents. Finding this posterior distribution for different kinds of data and models is a central problem in Bayesian statistics. See Bernardo and Smith [1994] and Gelman et al. [1995] for general introductions to Bayesian statistics.

In our nonparametric setting, we must find a posterior distribution on countably infinite collections of objects—hierarchies, path assignments, and level allocations of words—given a collection of documents. Moreover, we need to be able to do this using the finite resources of the computer. Not surprisingly, the posterior distribution for hLDA is not available in closed form. We must appeal to an approximation.

We develop a Markov chain Monte Carlo (MCMC) algorithm to approximate the posterior for hLDA. In MCMC, one samples from a target distribution on a set of variables by constructing a Markov chain that has the target distribution as its stationary distribution [Robert and Casella 2004]. One then samples from the chain for sufficiently long that it approaches the target, collects the sampled states thereafter, and uses those collected states to estimate the target. This approach is particularly straightforward to apply to latent variable models, where we take the state space of the Markov chain to be the set of values that the latent variables can take on, and the target distribution is the conditional distribution of these latent variables given the observed data.

The particular MCMC algorithm that we present in this paper is a Gibbs sampling algorithm [Geman and Geman 1984;

Figure 2 Screen shot for BOOKMARK

4. Research questions and hypotheses

In general terms, we attempted to investigate whether difference exists between THC-DAT and BOOKMARK in reading, specifically whether THC-DAT was more effective than BOOKMARK. Beyond that, we intended to measure user subjective evaluation to the two competing tools. Specific measures are described in Section 5.4.

More formally, several hypotheses were formulated on the basis of the expected performance of, and user subjective evaluation to THC-DAT and BOOKMARK. These hypotheses are with justifications:

H1: THC-DAT is more efficient than BOOKMARK

In THC-DAT, paragraphs scattered in different places in a long document are associated by a certain topic, so users can read relevant paragraphs directly, rather than spend time navigating full text to find relevant paragraphs and reading irrelevant paragraphs. Therefore, we assume that the time to complete a question-answering task is less using THC-DAT compared with BOOKMARK.

H2: THC-DAT is more effective than BOOKMARK

Compared with coarse-grained perspective, THC-DAT extracts hierarchical topics in a fine-grained way. Besides, with the topic tree, users can exploringly learn about unfamiliar topics with little cognitive burden and deepen understanding of a document. Therefore, we assume that users can complete tasks more effectively using THC-DAT compared with BOOKMARK.

H3: Users will think THC-DAT is more useful than BOOKMARK.

H4: Users will think THC-DAT is more usable than BOOKMARK.

H5: Users will prefer to use THC-DAT compared with BOOKMARK.

H6: Users will be more satisfied when using THC-DAT compared with BOOKMARK.

Hypotheses refer to user subjective evaluation are based on the hypothesized efficiency and effectiveness of THC-DAT. Furthermore, we believe that THC-DAT will provide a better user experience when performing the question-answering task.

5. Experiment design

5.1 Research method and data set

Control experiment was applied as the overall data collection method. Every participant was required to use the two tools to finish question-answering tasks about a document. The document used in our experiment was "The Nested Chinese Restaurant Process and Bayesian Nonparametric inference of Topic Hierarchies" authored by Blei. It talks about topic model (statistic algorithms for discovering the latent meaningful information) and discusses how the nested Chinese restaurant process (nCRP) which is one kind of stochastic process (by simulating the probability if any customer belonging to any table) is used in a Bayesian nonparametric statistics to build topic hierarchy tree. In order to maintain consistency, the same document was used in the two systems.

The aim of the experiment is to compare different reading consequences along several task dimensions, including: locating specific information from within the document, extracting gist of paragraphs, making global comprehension of the document. So experiment data was collected to compare the two systems on the following 3 types of tasks:

fact finding - requires participants to find out obvious and direct factual information from the document;

partial understanding - requires participants to understand partial content of the document;

full-text understanding - requires participants to analyze the full text in-depth and understand its theme, context and content.

full-text understanding task was presumed to be the most difficult type, followed by partial understanding and fact finding task, on the basis of the amount of information participants had to access to perform the task. Each type of task has two questions inside. The participants were given a total of 6 questions each (2 of each of the 3 task types). For half the questions they used the THC-DAT, and the control system for the other half. And all questions can also be categorized into 3 scenarios which are keyword guiding scenario, theme extraction scenario and overall comprehension scenario. In keyword guiding scenario, questions' keyword(s) can be found directly in the navigation bar either of BOOKMARK or THC-DAT. For fact finding and partial understanding questions, their answer are related to keyword(s) or topic(s), so FF0, FF1, PU0 and PU1 are in keyword guiding scenario. Specifically, the keyword(s) of FF0 and PU1 can be found in the navigation bar of THC-DAT, and the keyword(s) of FF0, FF1 and PU0 can be found in the navigation bar of BOOKMARK. Since summarizing the theme of document is a typical task of full-text understanding, one of our full-text understanding questions, FU0, is set to find out the main topic of the document. So FU0 is in theme extraction scenario. In overall comprehension scenario, answering the question is more difficult than in other scenarios because the answer depends on logical understanding of full-text rather than certain keywords. FU1 is in such kind of scenario. In a word, we intend to evaluate whether the topics extracted by THC-DAT can help users in analyzing a multi-topic document in different scenarios.

During the experiment, three sets of questionnaires were sent to the participants. The pre-experiment questionnaire was to collect the demographic information. The post-system questionnaire referred to users' overall evaluation of the system which conducted twice after participants finished questions on each system. Meanwhile, the answer and time spent on each question were collected through experimental logs.

5.2 Procedures

To eliminate learning and sequencing effect, the combination of users was randomized, tasks and systems using Latin Square so that there were 12 different sequences of tasks and systems. 36 participants were recruited, and thus each sequence was performed three times. The experiment design scheme is shown in Table 1.

Table 1 Experiment design scheme [1]

Rotation Number	First task set Question(system)			Second task set Question(system)		
R1	FF0（B）	PU0（B）	FU0（B）	FF1（T）	PU1（T）	FU1（T）
R2	PU0（B）	FU0（B）	FF1（B）	PU1（T）	FU1（T）	FF0（T）

[1] BOOKMARK is 'B', and THC-DAT is 'T', 'FF' represents fact finding task, 'PU' represents partial understanding task, 'FU' represents full-text understanding task. Questions within each type of task were marked as '0' and '1' in order.

R3	FU0（B）	FF1（B）	PU1（B）	FU1（T）	FF0（T）	PU0（T）
R4	FF1（B）	PU1（B）	FU1（B）	FF0（T）	PU0（T）	FU0（T）
R5	PU1（B）	FU1（B）	FF0（B）	PU0（T）	FU0（T）	FF1（T）
R6	FU1（B）	FF0（B）	PU0（B）	FU0（T）	FF1（T）	PU1（T）
R7	FF0（T）	PU0（T）	FU0（T）	FF1（B）	PU1（B）	FU1（B）
R8	PU0（T）	FU0（T）	FF1（T）	PU1（B）	FU1（B）	FF0（B）
R9	FU0（T）	FF1（T）	PU1（T）	FU1（B）	FF0（B）	PU0（B）
R10	FF1（T）	PU1（T）	FU1（T）	FF0（B）	PU0（B）	FU0（B）
R11	PU1（T）	FU1（T）	FF0（T）	PU0（B）	FU0（B）	FF1（B）
R12	FU1（T）	FF0（T）	PU0（T）	FU0（B）	FF1（B）	PU1（B）

To avoid the interaction between participants, the experiment was ran with each participant individually. Every participant was asked to complete the question-answering tasks as possible as they can. If they couldn’t get the answers indeed, they may submit “I can’t”, and no time limited for the question answer. Since participants were familiar with the BOOKMARK system, more training was focused on the THC-DAT system in which participants were asked to finish a training task individually to learn the function of the THC-DAT system deeply. The whole experiment steps are shown below.

(1)filling in the pre-experiment questionnaire;

(2)briefing on the experiment goal, and training on the BOOKMARK system and the THC-DAT system;

(3)finishing the training task on the THC-DAT system individually;

(4)using the first system to finish three question-answering tasks;

(5)filling in the post-system questionnaire for the first system;

(6)using the second system to finish three question-answering tasks;

(7)filling in the post-system questionnaire for the second system.

5.3 Participants

The 36 participants were all graduate students drawn from various schools of Wuhan University randomly to exclude major difference effects. Though the document used in our experiment was relatively difficult and long in English, all participants for the study had passed the College English Test Level 6 and have English reading and cognitive ability. Based on a pre-experiment questionnaire, summary statistics about the participants are presented in Section 6.1.

5.4 Measures

Objective and subjective measures are made to compare the performance of the two systems. Our objective measures were derived from the log data: answer to each question and time for each question. The form of each question is either multiple choice or quiz. For multiple choice, the score is either 0 or 1, with 0 being absolutely wrong and 1 being absolutely right. As for quiz, two experts judged the correctness of the answer with score ranged from 0 to 1 independently. Inter-rater reliability was assessed using Cohen's kappa (an inter-rater agreement measure), with 95% confidence intervals (Cohen, 1960; Gwet & Li, 2008). Using answer information, we were able to assess the effectiveness of task completion then reflect system effectiveness. The time for each question was recorded automatically in seconds. Using this information, we were able to assess the efficiency of task completion then reflect system efficiency.

Our subjective measures were extracted from user surveys from 4 aspects: usefulness, ease of use, intention to use and system satisfaction, and these were measured based on Davis's scale (1989) and Bhattacherjee's scale (2001). All subjective measures were rated on a scale of 1 to 5, where 1 means "strongly disagree" and 5 means "strongly agree". Furthermore, participants were asked to indicate the degree of familiarity with topic model before and after experiment for each system.

6. Experiment result analysis

6.1 Summary of participants data

Based on the pre-experiment questionnaire, summary data is presented on the participants. The basic demographics about the participants are as follows:

More female participants than male (61.1% vs 38.9%);

All participants ranged from 21 to 26 years old. The majority were first year and second year graduate students aged between 22-24. (41.7% were first year, 52.8% were second year, and 5.6% were third year);

The participants were from 16 different schools, covering humanities, social science, science and engineering. Specific information was shown in the following Table 2.

Table 2 The distribution of schools

Name of School	Frequency	Percent
School of Information Management	8	22.2
School of Computer Science	4	11.1
School of Mathematics and Statistics	3	8.3
School of Power and Mechanical Engineering	3	8.3
School of Economics and Management	3	8.3
School of Journalism and communication	3	8.3
School of Electronic Information	2	5.6
School of Marxism	2	5.6
School of Electrical Engineering	1	2.8
School of History	1	2.8
School of Life Science	1	2.8
School of Water Conservancy and Hydropower	1	2.8
School of Philosophy	1	2.8
School of Political Science and Public Administration	1	2.8
Research Institute of China's border and Marine	1	2.8
School of Resource and Environmental Science	1	2.8
Total	36	100.0

The participants assessed their familiarity with aspects of the question-answering task via the pre-experiment questionnaire as:

For the frequency of using electronic academic document, 83.3% of the participants used it more than 10 times per month, 13.9% were 3 to 10 times, the remaining 2.8% were less than 3 times;

For the frequency of using library database, 66.7% used it more than 10 times per month, 22.2% were 3 to 10 times, the remaining 11.1% were less than 3 times;

In terms of the computer application ability, 2.8% could only chat on the Internet expertly, 33.3% could also operate a variety of office software expertly, 30.6% could install a variety of systems and software, in addition to the above abilities. The remaining 33.3% could use at least one computer language for programming expertly;

As for the understanding of topic model, only 11.1% had a little understanding of it, the remaining were totally or basically unfamiliar with it.

To summarize, the participants in the experiment were more female, whose majors were widely distributed in multiple disciplines and schools, and were mainly first year and second year graduate students aged between 22-24. Most of them used electronic academic document and library database frequently, and had a high standard of computer application ability. But they knew little about topic model.

6.2 Analysis of objective measures

In this section, we focus on the presentation and analysis of the objective data which were relevant to system efficiency, as measured by 'time for task', and system effectiveness, as measured by 'score for task'. The average value of time and answer scores of each system were calculated over all questions and participants.

For each of the variables, time for task, score for task, an analysis of variance (ANOVA) was conducted to assess the significance of the two factors: 'system' (BOOKMARK or THC-DAT) and 'task type' (fact finding, partial understanding or full-text understanding).

Table 3 Summary statistics for comparisons between BOOKMARK and THC-DAT system

	Mean (standard deviation)		p-value
	BOOKMARK	THC-DAT	
Time	307.309 (245.478)	298.843(243.267)	0.799
Score	0.602(0.474)	0.762 (0.401)	0.008

In Table 3, for the efficiency measured by time for task, participants completed all three tasks without significantly faster (p=0.799) using THC-DAT (M=298.843s, SD=243.267s) compared with BOOKMARK (M=307.309s, SD=245.478s). Although there is no significant efficiency between the two systems, the mean time that users spent on THC-DAT is less than on BOOKMARK. That means hypothesis H1 is not strongly supported. Moreover THC-DAT is at least not inferior to BOOKMARK in efficiency and improvement is still needed in THC-DAT, the mean improvement is 8.466s.

In terms of score measure of effectiveness, the difference between two systems is very significant at level p=0.008. With participants scoring higher to complete a set of tasks with THC-DAT (M=0.762, SD=0.401) than with BOOKMARK (M=0.602, SD=0.474), there is very strong evidence to support the hypothesis H2 and conclude that THC-DAT is more effective than BOOKMARK.

In general, THC-DAT shows better effectiveness than BOOKMARK, and for efficiency, THC-DAT is not inferior to BOOKMARK at least. Furthermore, the efficiency and effectiveness in each kind of task are needed to analyze for THC-DAT's evaluation.

Table 4 Summary statistics for comparisons among three types of task

	Mean (standard deviation)			p-value
	fact finding	partial understanding	full-text understanding	
Time	265.948(190.182)	331.117 (265.743)	312.164 (266.432)	0.258
Score	0.792 (0.409)	0.781 (0.336)	0.472 (0.503)	0.000

The results in Table 4 showed the scores that participants got are statistical different among different types of tasks (p=0.000). The mean and standard derivation of scores for fact finding task, partial understanding task and full-text understanding task are (M=0.792, SD=0.409), (M=0.781, SD=0.336), (M=0.472, SD=0.503) respectively. The mean score of fact finding is the highest, partial understanding is the second, and both are higher than full-text understanding. Note that full text understanding questions are the most difficult, partial understanding questions are moderate, fact finding questions are the easiest. So, the order of scores for each kind of task seems to be expected. However, the time that participants spent are not statistical significant among different types of tasks (p=0.258). The mean and standard derivation of time for fact finding task, partial understanding task and full-text understanding task are (M=265.948s, SD=190.182s), (M=331.117s, SD=265.743s), (M=312.164s, SD=266.432s) respectively. The time that participants spent on each kind of task, in ascending order, is fact finding task, full-text understanding task and partial understanding task. The mean time of partial understanding is the highest, fact finding is the lowest, and it is very interesting that the time used for the full-text understanding task, which is the most difficult type, even less than for partial understanding task which is of medium difficulty.

Table 5 Summary statistics of time and score by system and question-answering task type

	Type of Task	Mean		Standard deviation		p-value
		B	T	B	T	
time	FF	267.613	264.284	184.599	198.215	0.941
	PU	386.206	276.027	335.709	155.926	0.078
	FU	268.109	356.219	166.856	334.912	0.162
score	FF	0.694	0.889	0.467	0.319	0.043
	PU	0.722	0.840	0.391	0.262	0.137
	FU	0.389	0.556	0.494	0.504	0.161

In order to clarify the difference between the two tools in dealing with each kind of task, specific data are reported in Table 5. It can be observed that only the score for fact finding task that participants got are significantly higher (p=0.043) on THC-DAT (M=0.889, SD=0.319) compared with BOOKMARK (M=0.694, SD=0.467), and users spent similar time using THC-DAT (M=264.284s, SD=198.215s) and BOOKMARK (M=267.613s, SD=184.599s), with p=0.941. For partial understanding task, time spent on it using THC-DAT (M=276.027s, SD=155.926s) is much less than BOOKMARK (M=386.206s, SD=335.709s), the difference is nearly significant (p=0.078). And participants still get higher score on THC-DAT (M=0.840, SD=0.262) than BOOKMARK (M=0.722, SD=0.391). For full-text understanding task, time spent on this kind of task using BOOKMARK (M=356.219s, SD=334.912s) reduced dramatically compared with the partial understanding task, while increased using THC-DAT (M=356.219s, SD=334.912s), and there is no discernible difference between the two systems (p=0.162). The mean and standard derivation of scores for THC-DAT and BOOKMARK are (M=0.556, SD=0.504), (M=0.389, SD=0.494) respectively, with p=0.161. This shows the effectiveness of THC-DAT is better than BOOKMARK for full-text understanding task.

In Figure3, the means of time on THC-DAT increase along with the increasing of task difficulty. However, the result on BOOKMARK shows that when task's difficulty increased, the mean time on BOOKMARK shows inverted U shaped curve. It increased first and then decreased. And Figure 4 shows that for each kind of task, users scored better using THC-DAT than BOOKMARK.

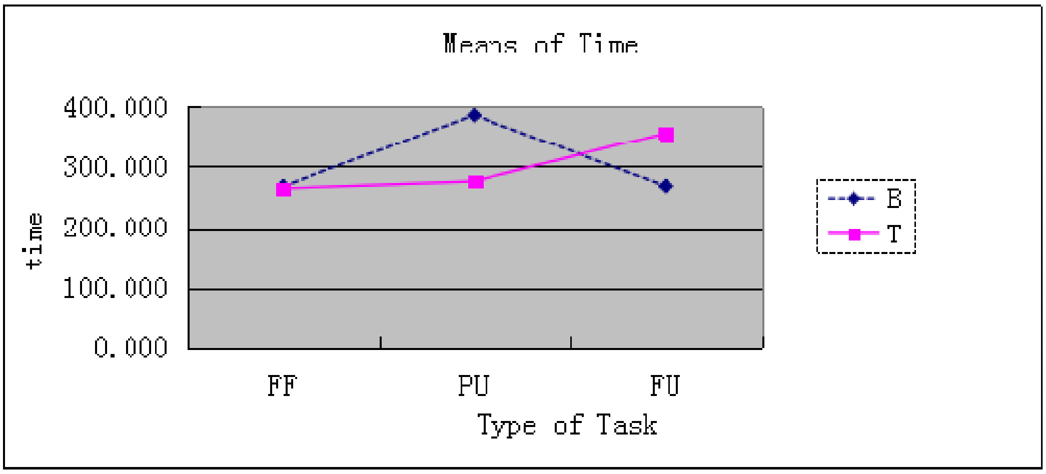


Figure 3 Mean time by system and question-answering task type

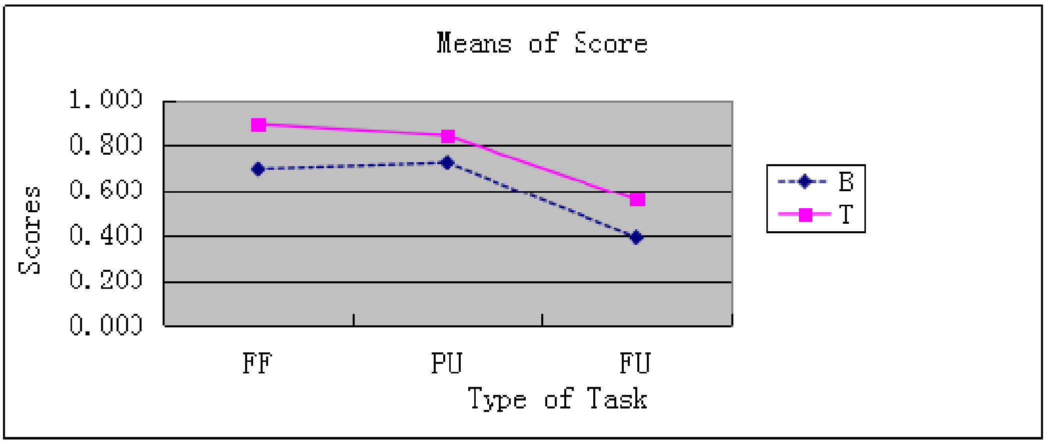


Figure 4 Mean score by system and question-answering task type

A more in depth results of the two systems in processing specific questions are displayed in Table 6. Only PU1 was completed significantly faster (p=0.024) on THC-DAT (M=256.977s, SD=139.155s) compared with BOOKMARK (M=430.526s, SD=279.786s). With respect to the score, the score for FF0 that participants got on THC-DAT (M=0.778, SD=0.428) was considerably higher (p=0.041) than on BOOKMARK (M=0.444, SD=0.511). Users got better scores for the rest questions using THC-DAT than BOOKMARK although there were no significant differences.

Table 6 Mean time and score by system and question-answering task

	Task	Mean(Standard deviation)		p-value
		B	T	
time	FF0	373.253(181.272)	261.896(182.055)	0.075
	FF1	161.973(116.903)	266.671(218.479)	0.082
	PU0	341.886(386.770)	295.078(172.983)	0.642
	PU1	430.526(279.786)	256.977(139.155)	0.024

	FU0	175.270(81.698)	180.756(96.101)	0.855
	FU1	360.948(179.981)	531.683(395.598)	0.105
score	FF0	0.444(0.511)	0.778(0.428)	0.041
	FF1	0.944(0.236)	1.000(0.000)	0.324
	PU0	0.889(0.323)	0.944(0.236)	0.560
	PU1	0.556(0.389)	0.736(0.250)	0.106
	FU0	0.389(0.502)	0.667(0.485)	0.100
	FU1	0.389(0.502)	0.444(0.511)	0.744

6.3 Analysis of subjective judgment

Here we present subjective judgment results obtained from the post-system questionnaires filled in by the participants after they finished each task set. See Appendix A for all questionnaires details.

In Table 7, the mean, median, standard deviation of the responses (5-point scale) are given for each system for questions about usefulness, ease of use, intention to use and system satisfaction from post-system questionnaires. The mean values of several questions' responses corresponding to each indicator were calculated for the individual participant, and then averaged over participants. Two systems are compared based on these averaged data using paired t test.

In relation to questions on ease of use, participants rate THC-DAT (M=4.16, Med=4.08, SD=0.451) as not that easier to use in performing these tasks than BOOKMARK (M=4.10, Med=4.00, SD=0.437), and there is no discernible difference between two systems (p=0.431). This conclusion keeps consistent with our objective results of time for task in Table 3 where the efficiency is not significant between the two systems.

In the process of answering questions, participants think THC-DAT (M=4.04, Med=4.00, SD=0.497) is more useful than BOOKMARK (M=3.57, Med=3.67, SD=0.746), with a significant level p=0.001. This conclusion shows coincide with score for task in Table 3 where participants obtain higher score on THC-DAT than on BOOKMARK.

The intention to use was assessed by participants as prefer to use THC-DAT (M=3.57, Med=3.67, SD=0.746) than BOOKMARK (M=3.57, Med=3.67, SD=0.746), and this is significant at p=0.014.

In terms of system satisfaction, participants are more satisfied with THC-DAT (M=3.57, Med=3.67, SD=0.746) than BOOKMARK (M=3.57, Med=3.67, SD=0.746), and this is significant at p=0.010.

On the basis of subjective results, there is very strong evidence to support the hypothesis H3, H5 and H6, while H4 is inconclusive.

Table 7 Subjective judgment for comparisons between BOOKMARK and THC-DAT system

	Mean		Median		Standard deviation		p-value
	B	T	B	T	B	T	
Usefulness	3.57	4.04	3.67	4.00	0.746	0.497	0.001
Ease of use	4.10	4.16	4.00	4.08	0.437	0.451	0.431
intention to use	3.59	3.92	3.67	4.00	0.731	0.634	0.014
system satisfaction	3.56	3.87	3.75	4.00	0.733	0.575	0.010

In addition, recall that we asked participants to indicate the degree of familiarity with topic model before and after experiment for each system from the summary of participants data, we know that participants knew little about topic model before the experiment. In contrast, participants reflected they got more familiar with topic model after using THC-DAT (M=2.97, Med=3.00) than BOOKMARK (M=2.86, Med=3.00), which also demonstrates that THC-DAT is more effective than BOOKMARK from another perspective.

7. Discussion

The results show that THC-DAT is more helpful than BOOKMARK. On the one hand, the mean score of all tasks on THC-DAT is higher than BOOKMARK over 26.6%, with a significant difference. Especially for question FF0, THC-DAT shows statistical edge over BOOKMARK with p=0.041. On the other hand, in relation to the efficiency concerning completion time, there is no evidence for significant difference between the two tools, but the mean time spent on THC-DAT is less than on BOOKMARK, and THC-DAT achieves statistical advantage in question PU1 with p=0.024 (See Table 3 and Table 6). So generally, THC-DAT manifests more effective than BOOKMARK for the experimental tasks; meanwhile, THC-DAT is at least not inferior to BOOKMARK in efficiency.

Since no considerable difference appeared in efficiency, maybe the time spent on each kind of task using two tools can help us to explain the result. Note that the time used for the full-text understanding task on BOOKMARK experienced a dramatic decline, with average time 118.097s less than partial understanding task and the whole time curve on BOOKMARK shows inverted U shape. However, the score is lowest in full-text understanding task accordingly. In another word, the time dropped rapidly and the scores became worse as well on BOOKMARK (See Table 5, Figure 3 and 4). In contrast, on THC-DAT, the time users spent increased with the increasing of task difficulty but the score they got decreased, which is always higher than that they got on BOOKMARK. This may reflect the different effects on cognitive load between THC-DAT and BOOKMARK. Previous researchers have verified there are two stages of cognitive load evolving along with the increasing of task difficulty. In the first stage, higher difficulty of task leads to higher cognitive loads (Kahneman & Beatty, 1966), such as more time users spent and lower score users got (Johnston, Fiore & Smith, 2013; O'Donnell & Eggemeier, 1986). Meanwhile, in the second stage, Granholm, Asarnow & Sarkin's research (1996) demonstrated that people in face of very difficult cognitive tasks will reach the state of cognitive overload and the time they cost will reduce. Moreover, Marshall's (2002) research showed that user experiencing higher cognitive load will be more likely to make an error while performing a task. Based on the studies of Granholm *et al.* and Marshall, it can be concluded that the time users cost becomes less and the score they got become worse when they reaching the state of cognitive overload. In this research, note that full-text understanding task is the most difficult, and users cost less time while scored worse on BOOKMARK, representing the second stage, so full-text understanding task leads to cognitive overload on BOOKMARK according to the above conclusion. However, users cost more time and scored worse on THC-DAT for full-text understanding task, representing the first stage before the state of cognitive overload. This also means THC-DAT can slow down the process of approaching cognitive overload while BOOKMARK can't.

As we presented in Section 5.1, questions reflect different task types and different scenarios. In order to identify in which aspects THC-DAT is more effective or efficient than BOOKMARK, the results of each question in different scenario are analyzed further.

First, in keyword guiding scenario, all questions are fact finding task or partial understanding task. THC-DAT's navigation bar contains keywords of questions FF0 and PU1, and shows certain superiority in time and score (See Table 6). The time spent on FF0 using THC-DAT is less than BOOKMARK with an approximately significant difference (p=0.075), and there is considerable difference in score at p=0.041. As for PU1, users completed it significantly faster on THC-DAT (p=0.024) and the mean score on THC-DAT is higher than BOOKMARK over 32%. These show that THC-DAT is perfect for keyword guiding scenario when it can extract the topic term of question. Moreover, it is notable that when users complete FF0, its keyword(s) appeared in the navigation bar of both tools, while PU1's keyword(s) can only be found in THC-DAT, which indicates that difference exists in keyword guiding capability between the two tools. Since THC-DAT extracts more topic terms and organizes the hierarchical relationship of contents between topics effectively, contents that THC-DAT guides are more concrete and the content boundaries of each topic is more accurate.

Second, in regard to the remaining questions in keyword guiding scenario, namely FF1 and PU0, their keyword(s) were displayed in the navigation bar of BOOKMARK but not in THC-DAT. Although there is no statistically significant difference in time and score between the two tools, the results in Table 6 reveal that THC-DAT is superior to BOOKMARK. For FF1, BOOKMARK achieves slightly faster in time with p=0.082, and the score is worth thinking. All answers are right on THC-DAT but not all on BOOKMARK. Similarly, although no significant differences appeared in time and score for PU0, users spent less mean time on THC-DAT and scored higher. So, even if THC-DAT cannot extract all topics of the document, but, because the topics it extracted contain multiple terms and are hierarchically organized with their contents and this revealed some semantic relations between topics which contributes to narrowing users' search scope. It makes sense that THC-DAT can still deal with questions like FF1, PU0 well, In other words, THC-DAT also fit keyword guiding scenario when it cannot extract the topic term of question.

Through above analysis, we find when THC-DAT extracts evident topic terms, it outperforms BOOKMARK obviously. When explicit terms exist in BOOKMARK but not in THC-DAT, THC-DAT still has an advantage in score. This is mainly because the function of THC-DAT is extracting multiple topic terms within a document and then expressing logical relationships between topics hierarchically. Though THC-DAT cannot extract all keywords in questions occasionally, its multiple terms and hierarchical relations between topics do assist users effectively. In addition, fact finding is defined as acquiring detailed knowledge about the factual information (Conflict Research Consortium, 1998). Partial understanding means one seeks to grasp at least a partial reason for a pattern of how things work or why a phenomenon occurs (Keil, 2011). In light of these definitions, it is obvious that dealing with fact finding and partial understanding problems in academic papers often involves multi-grained topics and information processing, so we deduce that THC-DAT is helpful in keyword guiding scenario.

Additionally, as for question FU0 which represents the theme extraction scenario, the completion time on THC-DAT is pretty close to that on BOOKMARK, while users scored significantly higher (p=0.100) on THC-DAT than on BOOKMARK, with a difference of over 71%. The result still ascribes to THC-DAT's core function, : extracting topics hierarchically, namely topic more general will correspond to the root in the hierarchical tree while topic more concrete will correspond to the leaf. Note that FU0 was completed so fast on BOOKMARK that it was only slower than fact-finding questions, and the corresponding score is the lowest. Based on relevant cognitive overload theory (Granholm *et al.*, 1996; Kahneman & Beatty, 1966), we infer that users achieved the state of overload when answering FU0 on BOOKMARK. On the contrary, users can obtain the theme of a document quickly from the root topic in the hierarchical tree on THC-DAT, and get relatively higher score than that on BOOKMARK. Therefore, it is clear that THC-DAT is especially suitable for the theme extraction scenario, because users can acquire full-text theme accurately on THC-DAT and avoid achieving the state of overload effectively.

As for question FU1 , which represents the overall comprehension scenario, since answering this question cannot merely rely on any keywords, FU1 turns out to be the most difficult question. On BOOKMARK, the time and score are both worth discussing. The completion time is even less than fact-finding question FF0 and partial-understanding question PU1, the score is the lowest as well, which continues the cognitive overload state of FU0. On THC-DAT, users spent the longest time dealing with FU1 though scored worse. Combining with cognitive overload theory (Granholm *et al.*, 1996; Kahneman and Beatty, 1966), it can be inferred from the similar scores and different mean time that, even if the score seems low on both tools, users tend to give up on BOOKMARK but stick to completing the task on THC-DAT when facing with demanding questions, which further reflects that THC-DAT can improve users' willingness to undertake difficult task and increase the expectancy of fulfilling the task.

The results of subjective evaluation show that THC-DAT is highly acknowledged in usefulness, intention to use and system satisfaction than BOOKMARK. There is no discernible difference in ease of use and their mean values are close to each other (See Table 7), but BOOKMARK has been widely used and users are familiar with it, while THC-DAT is a new tool, this demonstrates that THC-DAT is not inferior to BOOKMARK in ease of use. In a word, THC-DAT surpasses BOOKMARK in general satisfaction.

In conclusion, practical and theoretical implications of this study can be drawn as follows:

On one hand, applying THC-DAT to digital libraries or electrical document reading systems would enhance users' reading performance, willingness to undertake difficult task and satisfaction. First, BOOKMARK is a common assistant tool in digital libraries and electrical document reading systems, so research and improvement on it is of important promotion value. Since THC-DAT is superior to BOOKMARK in effectiveness and not inferior in efficiency for a variety of tasks and scenario types, it is suggested to adopt THC-DAT to promote users' reading effects in digital libraries and electrical document reading systems. Especially, when analyzing document theme or the topics in THC-DAT covering questions' keywords, THC-DAT shows assistant effects resemble state of art. Since THC-DAT extracts many topic terms and organizes hierarchical relations effectively, its ideal performance can be sure to occur frequently.

Second, THC-DAT could help document analysis researches to gather more user data. Many document analysis techniques rely on mining user-involved data stemmed from user-generated content or user behavior log, such as folksonomy, user preference learning and evaluating applications (Agosti, Crivellari & Nunzio, 2011; Egger & Lang, 2012; Francisco, Ricardo & Oliveira, 2012; Lin, Huang & Chen, 1999). However, the difficulty to complete a task has a negative impact on the willingness of participants (Smith, 1999). When users have lower participation willingness, consequently the user-involved data will be generated less, and this may hinder the practical use and evaluation of the document analysis technique which based on the user-involved data. We confirmed that THC-DAT can improve users' willingness to undertake difficult task in overall comprehension scenario, hence, applying THC-DAT to these techniques contributes to improving users' willingness to undertake difficult task as well, namely generate rich basic data.

Additionally, BOOKMARK is widely used but may not be a satisfying tool for users to read within documents. It is found that THC-DAT is highly acknowledged by users in usefulness, intention to use and system satisfaction than BOOKMARK. So it is recommended to adopt THC-DAT in electrical document auxiliary reading and library knowledge organization service to improve users' satisfaction.

On the other hand, THC-DAT is conducive to addressing the cognitive overload problems in current research domain and applications, such as document relevance judgment or indexing in information retrieval based on relevance feedback technique. These researches and applications often involve users to judge the theme of document or different parts. Our study of extracting full-text theme on BOOKMARK indicates that this task type might increase cognitive load and users are inclined to reach cognitive overload, which will affect the accuracy and validity of relevance feedback technique. So it is suggested that, cognitive overload should be considered in information retrieval studies based on relevance feedback technique. In contrast, THC-DAT is characterized by hierarchical topic extraction, and thus it enables users to extract document theme in root of the hierarchical topic tree accurately, alleviate cognitive load and amplify cognitive capability, so THC-DAT aids users in analyzing full-text theme quickly, judging document relevance and indexing document better. This makes it a good choice to avoid cognitive overload in relevance feedback. Moreover, because of its state-of-art performance in extracting document theme and hierarchical semantic structure in multi-topic documents, THC-DAT is of great value to researches and applications related to document topics, such as document visualization and document summarization.

8. Conclusion

In this study, we conducted a controlled experiment to evaluate the effectiveness, efficiency and user satisfaction of THC-DAT, a multi-topic document analysis tool. Our experiment results demonstrate that THC-DAT performs better in reading performance than the benchmark system, BOOKMARK, in terms of effectiveness. Meanwhile, THC-DAT manifests not inferior to BOOKMARK in efficiency. It was also evaluated more favorably than BOOKMARK on several perceptual dimensions (i.e., usefulness, intention to use, system satisfaction). Though no significant difference appeared in ease of use. However, in general, users appreciated THC-DAT more. In the discussion, we did a more detailed analysis for specific questions in different scenarios. Since THC-DAT extracted topics hierarchically and displayed their logical relations, it outperforms BOOKMARK in reading assistance in keyword guiding scenario, alleviate users' cognitive load in theme extraction scenario, and improve users' willingness to undertake difficult task in overall comprehension scenario. Two implications in terms of theoretical and practical are concluded based on the analysis above. First, THC-DAT can be applied to digital library or electrical document reading system to enhance users' reading performance, willingness to undertake difficult task, and satisfaction. Second, as for theoretical implication, THC-DAT represents an example to address cognitive overload problem in the domain of information retrieval.

9. Acknowledgments

The authors gratefully acknowledge the financial support for this work provided by National Natural Science Foundation of China (No:71303089, 71273195 and 71420107026), the National Basic Research Program of China (973 Program, No: 904171200) and the Major projects of the National Social Science Foundation. (No:13&ZD183)

References

Agosti, M., Crivellari, F., & Nunzio, G. M. D. (2012). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Mining & Knowledge Discovery, 24(3), 663-696.

Bhattacherjee, A. (2015). Understanding information systems continuance: an expectation-confirmation model. Mis Quarterly, 25(3), 351-370.

Byrd, D. (1999). A scrollbar-based visualization for document navigation. Computer Science, 122-129.

Cao, N. and Wang, D. (2000), Research on content information description of Chinese book, Journal of The National Library of China , 9(2), 26-31 (in Chinese).

Casey, M. E., & Savastinuk, L. C. (2006). Library 2.0: service for the next-generation library. Library Journal, 131, 3.

Chen, H. Y., & Si, L. (2011). Construction of the mode of the user-involved library information organization under web2.0. Information & Documentation Services, 32(3), 62-66 (in Chinese).

Chen, J., Wang, T. T., & Lu, Q. (2016). THC-DAT: a document analysis tool based on topic hierarchy and context information. Library Hi Tech, 34(1), 64-86.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational & Psychological Measurement, 20(1), 37-46.

Conflict Research Consortium (1998). "General Information on Fact-Finding", available at: http://www.colorado.edu/conflict/peace/problem/factfinding.htm (accessed July 8, 2016).

Dang, Y., Zhang, Y., Chen, H., Brown, S., Hu, P., & Nunamaker, J. (2012). Theory-informed design and evaluation of an advanced search and knowledge mapping system in nanotechnology. Journal of Management Information Systems, 28(4), 99-128.

Davis, Fred D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. Management Information Systems Quarterly, MIS Quarterly,13 (3), 319-339.

Egger, M., & Lang, A. (2012). A brief tutorial on how to extract information from user-generated content (ugc). KI - Künstliche Intelligenz, 27(1), 53-60.

Francisco, A. P., Ricardo, B., & Oliveira, A. L. (2012). Mining query log graphs towards a query folksonomy. Concurrency & Computation Practice & Experience, 24(17), 2179-2192.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. Diseases of the Colon & Rectum, 33(4), 1475-1479.

Gretarsson, B., Donovan, J., Bostandjiev, S., Llerer, T., Asuncion, A., & Newman, D., et al. (2012). Topicnets: visual analysis of large text corpora with topic modeling. Acm Transactions on Intelligent Systems & Technology, 3(2), 565-582.

Gwet, & Li, K. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical & Statistical Psychology, 61(Pt 1), 29-48.

Harper, D. J., Koychev, I., Sun, Y., & Pirie, I. (2004). Within-document retrieval: a user-centred evaluation of relevance profiling. Information Retrieval, 7(3-4), 265-290.

Hersh, W., Pentecost, J., & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science, 47(1), 50 – 56.

Johnston, J. H., Fiore, S. M., Smith, C. A. P., & Orlando, N. (2013). Application of cognitive load theory to develop a measure of team cognitive efficiency. Military Psychology, 25(3), 252-265.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. Science, 154(3756), 1583-1585.

Keil, F. (2011). The problem of partial understanding. Current Trends in LSP Research: Aims and Methods Series: Linguistic Insights, 44, 251-276.

Lemarié, J., Eyrolle, H., & Cellier, J. M. (2006). Visual signals in text comprehension: how to restore them when oralizing a text via a speech synthesis?. Computers in Human Behavior, 22(6), 1096-1115.

Lin, I. Y., Huang, X. M., & Chen, M. S. (1999). Capturing user access patterns in the web for data mining. 345-348.

Liu, Y. (2003). The textualisation of titles. Linguistics Study (in Chinese).

Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., & Cai, W., et al. (2009). Interactive, topic-based visual text summarization and analysis. Acm Transactions on Intelligent Systems & Technology, 3(2), 543-552.

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. Educational Psychology Review, 1(3), 209-234.

Lorch, R. F., Lorch, E. P., Ritchey, K., Mcgovern, L., & Coleman, D. (2001). Effects of headings on text summarization ☆. Contemporary Educational Psychology, 26(2), 171-191.

Maness, J. M. (2006). Library 2.0 theory: web 2.0 and its implications for libraries. Webology, 3(2).

Marshall, S. P. (2002). The Index of Cognitive Activity: Measuring cognitive workload. Human Factors and Power Plants, 2002. Proceedings of the 2002 IEEE, Conference on (pp.7-5-7-9).

Miller, N. E., Wong, P. C., Brewster, M., & Foote, H. (1998). TOPIC ISLANDS™ — A Wavelet-Based Text Visualization System. IEEE Computer Society.

Mizoguchi, K., Sakamoto, D., & Igarashi, T. (2013). Overview Scrollbar: A Scrollbar Showing an Entire Document as an Overview. Human-Computer Interaction – INTERACT 2013. Springer Berlin Heidelberg.

Nishihara, Y., Sato, K., & Sunayama, W. (2011). Text visualization using light and shadow based on topic relevance. Transactions of the Japanese Society for Artificial Intelligence, 2(6), 479-487.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology.

Schwartz, M., Hash, C., & Liebrock, L. (2010). Term distribution visualizations with focus+context: overview and usability evaluation. Multimedia Tools & Applications, volume 50, 509-532(24).

Smith, L. F. (1999). Difficulty, consequence, and effort in academic task performance.

Psychological Reports, 85(7), 869-879.

Spyridakis, J. H. (1989). Signaling effects: a review of the research-part i. Journal of Technical Writing & Communication, 19(3), 1-1.

Tagarelli, A., & Karypis, G. (2013). A segment-based approach to clustering multi-topic documents. Knowledge & Information Systems, 34(3), 563-595.

Wang, G. Z., Wen, C. K., Yan, B. H., Xie, C., Liang, R. H., & Chen, W. (2013). Topic hypergraph: hierarchical visualization of thematic structures in long documents. Sciece China Information Sciences, 56(5), 1-14.

Wu, D., He, D., & Xu, X. (2012). A study of relevance feedback techniques in interactive multilingual information access. Library Hi Tech, 30(3), 523-544.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., & Singhal, A. (2010). SCAN: Designing and Evaluating User Interfaces to Support Retrieval From Speech Archives. SIGIR '99: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, Ca, Usa (pp.26-33).

Yin, S. C. (2012). Heading syntax. Beijing: The Commercial Press (in Chinese).